**Hadoop assignment 2 – Pig** Job Vermeulen 616184

De gelopen stappen:

1. Start ‘Hortonworks docker sandbox’ op
2. Surf door middel van de browser naar 127.0.01:8080
3. Inloggen met maria\_dev en ww maria\_dev
4. Wacht tot alle services draaien
5. Ga naar files (naast de user settings dropdown)
6. In de files view, maak een map genaamd ***assignment2Files*** aan in ***/users/maria\_dev***
7. Upload de dataset
8. Ga nu naar de pig overview (naast de user settings dropdown)
9. Creëer een nieuw script voor assignment 2
10. Draai elk script met run on TEZ aangevinkt

Zie voor het geschreven script (hierin staat ook de uitleg van het script):

<https://github.com/Jobvermeulen/ParallelDistributedProcessing/tree/master/pdpAssignment2>

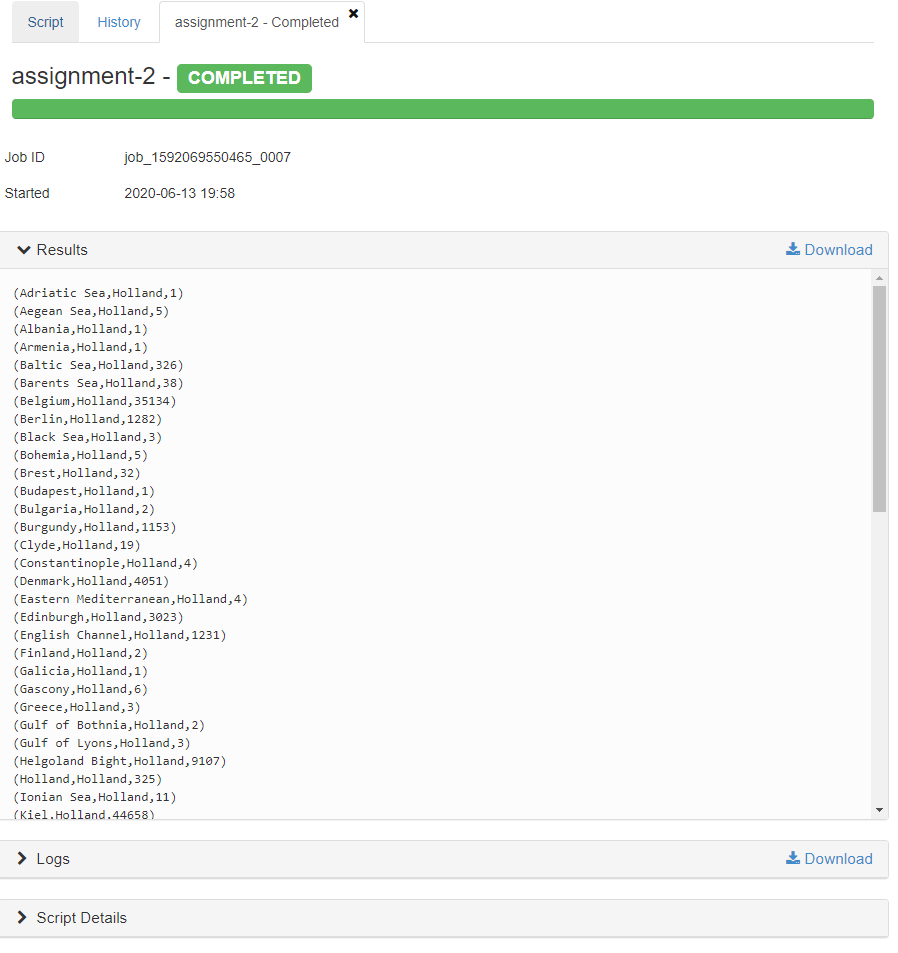
**Resultaat** – 6 opdracht:

De eerste stap bij deze opdracht was om de dataset op te halen. Dit werd gedaan door middel van een csv loader. Na het inladen van de data moeten de kolommen gedefinieerd worden.   
De data werd vervolgens gefilterd op de target, ‘Holland’.

De gefilterde dataset werd vervolgens gegroepeerd op locatie (‘Location’), daarnaast werd er voor iedere groep een tuple toegevoegd. Deze tuple bevat de volgende kolommen:

* De target
* De locatie
* De count van hoe vaak een locatie de target heeft aangevallen

Dit werd toegevoegd aan een resultaten variabelen. Deze variabelen kon vervolgens worden geordend op alfabetische volgorde. En als laatste werd het resultaat uitprint.

Dit is het resultaat daarvan:

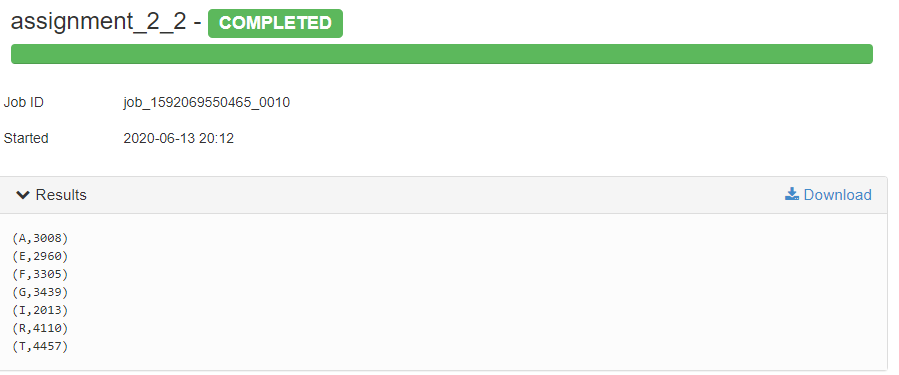
**Resultaat** – 7 opdracht:

De eerste stap bij deze opdracht was om de dataset op te halen. Dit werd gedaan door middel van een csv loader. Na het inladen van de data moeten de kolommen gedefinieerd worden.   
De data werd vervolgens gefilterd op de playersCSV, op won == 1.

Vervolgens werd er een groep aangemaakt waarin de gefilterde data werd gegroepeerd op country.

Op deze gegroepeerde data werden tuples gemaakt met de country, en de count hoe vaak er gewonnen is. Dit resultaat werd vervolgens geprint.

Zie hier het resultaat:



**Resultaat** – 8 opdracht:

De eerste stap is om de data in te laden. Dit gebeurd met de csv loader. Dit keer moeten er twee datasets ingeladen worden. De twee datasets worden vervolgens samengevoegd, dit gebeurd doormiddel van de id van de game dataset en de gameId in de players dataset.

De gecombineerde dataset wordt gefilterd op de winst. De gefilterde dataset wordt vervolgens gegroepeerd op country.

Met deze gegroepeerde dataset worden er tuples gemaakt. Deze bevatten de name van de country en een sum van de number of turns.

Deze tuples worden vervolgens aflopend geordend doormiddel van de number of turns. Hiervan wordt een selectie van de top 10 gemaakt.

Deze selectie wordt getoond.

Zie hier het resultaat:

